Validity of Colorado Learning Attitudes about Science Survey for a high-achieving, Finnish population

Inkeri Kontro*
Department of Physics, University of Helsinki, POB 64, FI-00014 Helsinki, Finland

David Buschhüter
Institut für Physik und Astronomie, Universität Potsdam, Karl-Liebknecht-Straße 24/25, 14476 Potsdam-Golm, Germany

(Received 14 April 2020; accepted 18 June 2020; published 14 July 2020)

The Colorado Learning Attitudes about Science Survey (CLASS) is an instrument which is widely used in physics education to characterize students' attitudes toward physics and learning physics and compare them with those of experts. While CLASS has been extensively validated for use in the context of higher education institutions in the United States, there has been less information about its use with European students. We have studied the structural, content, and substantive aspects of validity of CLASS by first doing a confirmatory factor analysis of $N = 642$ sets of student answers from the University of Helsinki, Finland. The students represented a culturally and demographically different subset of university physics students than in previous studies. The confirmatory factor analysis used a 3-factor, 15-item factor structure as a starting point and the resulting factor structure was similar to the original. Just minor modifications were needed for fit parameters to be in the acceptable range. We explored the differences by student interviews and consultation of experts. With the exception of one item, they supported the new 14-item, 3-factor structure. The results show that the interpretations made from CLASS results are mostly transferable, and CLASS remains a useful instrument for a wide variety of populations.

DOI: 10.1103/PhysRevPhysEducRes.16.020104

I. INTRODUCTION

Beginning physics students start with a set of expectations and beliefs about physics which often does not mirror those of practicing physicists. The development of these beliefs toward a more expertlike view is seen as an important learning outcome. In many cases, students even know what physicists would say in the situations probed, but do not agree with them [1]. In addition to being important in themselves, attitudes that align with experts also somewhat predict learning [2–5]. Instruments such as Views about Science Survey (VASS) [6], Maryland Expectations about Science Survey (MPEX) [7], and the Colorado Learning Attitudes about Science Survey (CLASS) [8] have been developed to measure these attitudes.

CLASS [8] is one of the most recent and widely used instruments for measuring attitudes. CLASS consists of 42 statements, which are scored by a five point Likert scale (a range of strongly agree to strongly disagree). The statements are formulated to be simple and concise, as well as usable in a variety of physics courses [8]. Unlike MPEX, the statements do not refer to studying on one course, but rather to broader views about (learning) physics. CLASS has been adapted for use in chemistry (CLASS-Chem) [9] and biology [10].

Data from different populations were used in the development of CLASS [8]. CLASS statements are intended to be concise and involve situations that, according to the authors, can arise in all kinds of physics studies. CLASS aims to cover both views about physics as a science and the practices and processes of learning physics. Hence, it has been used to study various kinds of student populations. In the development of CLASS, all questions were tested and validated by physicists and via student interviews. The structural validity was addressed by a factor analysis.

The original factor structure used 26 of the 41 questions divided into eight partly overlapping categories through so called reduced-basis factor analysis [8]. Douglas *et al.* [11] found evidence for three categories using fifteen statements by exploratory factor analysis. Cahill *et al.* [12] identified two categories for learning, using 25 statements, and proceeded to validate those with a confirmatory factor analysis (CFA). CLASS scores have been found to correlate weakly with learning outcomes [2,4,5] and experiencing high levels of challenge, interest, and skill at the same time (optimal learning moments) [13].

*inkeri.kontro@helsinki.fi

CLASS has also been translated into several languages, such as Chinese [14], Arabic [15], and Spanish [16]. When translating CLASS, much care has been taken so that the students understand the questions correctly. For example, the Chinese translation was validated by bilingual physicists and via student interviews [14]. The English version was also validated for a predominantly Hispanic sample by student interviews, which uncovered one misinterpretation. For an item which addressed a situation where the student does not remember an equation in an exam, many students said they would try to answer the question in other ways, whereas the expertlike reasoning would be to say that equations can be derived [17]. However, CLASS factor analyses have not been validated outside of United States (U.S.), nor are we aware of any "cross validations" inside the U.S. Despite this, the factors are used to make statements about students' attitudes toward physics without ascertaining that a specific set of statements (a factor) describes the same attitudes in a different population.

The categorizations of the CLASS questions hinge on two things: whether the statements group in a psychometric sense and whether the statements group with respect to these aspects from the perspective of the students. The first clause is less likely to have a cultural dimension that differs from country to country, as this categorization is done by expert physicists who may be expected to be a reasonably homogeneous population with regard to attitudes to good practices in learning physics. However, student populations differ significantly in, for example, age, gender distribution, major subject, and prospective careers depending on institution and country.

In lower education, such cultural differences are a closely studied topic. For example, the Organisation for Economic Co-operation and Development (OECD) administers the Programme for International Student Assessment (PISA), a worldwide study which assesses education systems. PISA includes a science survey, which involves both questions measuring the students' aptitude in science and their attitudes toward it, and the coupling between these varies by country. In Finland, learning outcomes and attitudes toward science, as measured by PISA, are more closely coupled than on average in OECD countries [18]. However, attitudes are not correlated with learning outcomes when results from all countries are pooled. In fact, attitudes of Finnish pupils are poorer than those of OECD peers, despite higher learning outcomes [18,19].

University physics students are naturally a very different selection than that of PISA, which aims to study the whole age group. At the University of Helsinki (UH), the population attending the physics courses is high achieving and very research oriented. In Finland, students study the basics of a wide range of subjects in high school, but specialize in a few, on which they write matriculation exams. Students who want to study physics further apply to physics or physical sciences bachelor programmes, and declare specialization (e.g., physics, theoretical physics, or astronomy) in their first or second year of studies. In university, science studies generally have a very strong focus on the declared subject and for science students, choosing minor subjects outside science is uncommon. Also, while many engineering students study physics, they are generally enrolled in technical universities and follow their own curriculum. Hence, the majority of students in physics courses are physics students, closely followed by other science students (including preservice math and physics teachers). There is thus a difference in student interest, the pre-university experience and the surrounding culture, compared to many large-enrollment physics courses in the U.S., and it is important to study the validity of CLASS for use in this student population.

Validity is a judgment of whether theoretical considerations and empirical evidence support the interpretations of test scores [20,21]. According to Messick, validity can be divided into six categories: the content aspect, substantive aspect, structural aspect, generalizability, the external aspect, and consequential aspect [20]. The content aspect of construct validity considers content relevance, representativeness and technical quality; the structural aspect relates the fidelity of the scoring to the underlying construct domain; the substantive aspect considers the theoretical rationales that explain the consistencies in response patterns; the generalizability aspect refers to the generalizability of results from one population to another, as well as the generalizability of the measured tasks to similar tasks; the external aspect to the correlations of test scores to external variables; and the consequential aspect to the consequences resulting from testing and result interpretation to the tested population [20].

We wanted to probe the different aspects of validity of CLASS interpretations for a group of students who are culturally and demographically different from U.S. college populations. Our hypothesis was that CLASS can be used to measure the expertlike attitudes of Finnish university students, and we expected the factor structure to be the same as found in other student populations. As the 3-factor model [11] is based on the assumption that three distinct constructs underlie student answers, this is the natural starting point of our analysis. Such constructs should be transferable to a different population, and this can be tested by a confirmatory factor analysis. We aimed primarily to validate the 3-factor model by confirmatory factor analysis and secondarily to adjust the obtained model if necessary with regard to fit indices.

However, the three factors in the model can never be observed directly opposed to the manifest responses themselves. One possible interpretation of latent trait models, such as linear factor models or probabilistic models like the Rasch model, is the realist view that the traits exist and cause the differences in observable test scores [22]. This in itself is a very bold hypothesis and

emphasizes the need for additional evidence from other perspectives than the structural view, as correlation between the item scores can have various reasons. Therefore we decided to collect evidence on the content and substantive aspects of validity as well. The study in itself can also be seen to evaluate the generalizability of CLASS for students representing different demographics.

II. METHODS

A. Quantitative data collection

The data were collected during the first introductory course in physics at the University of Helsinki during 2011–2016. The survey was given as one assignment in the weekly calculation exercises, and additional credit equal to one solved problem was awarded to all participants that completed the survey properly. The following criteria were used to discard improper sets of student data: Using less than 200 sec for finishing the survey, having more than 4 missing answers, having more than 18 same answers ($> 2/3$) in the 26 statements used in the factor analysis of Ref. [8], and an incorrect answer to the control item (31). In all years, the full CLASS (41 items) was collected, but for the validity analysis, only the data from questions used in the 3-factor model (15 items [11]) were used.

The survey was given in English in all years. Originally no validated translation into Finnish was available, and later no reason to change to the translation emerged. The course is instructed in Finnish, but materials used in introductory physics are in English, and the students are expected to have a working knowledge of English when starting their studies.

The data consist of 642 sets of student answers, of which 404 (63%) are physics majors, the rest being mainly other science (mathematics, chemistry, and geography) majors. 400 (62%) of the participants were male, 236 female, and 6 declined to say or chose "other." No information on student ethnicity was collected, as this student population is ethnically so homogeneous that collecting data on ethnicity may compromise the anonymity of minority students.

The sample of 642 students seemed appropriate for conducting a confirmatory factor analysis. There is no clear cutoff value for the sample size suitable for structural equation models or confirmatory factor analysis in general independent of the actual model [23]. A rule of thumb, however, is that $N:q > 20$ or > 10, where N is the number of cases (here 642 participants) and q are the model parameters to be estimated (here $23 = 16$ loadings $+4$ correlations $+3$ residual correlations, see Ref. [11]) [24,23]. In the case of this study $N:q$ satisfied both conditions ($N:q = 27.91$).

B. Expert rating of the 3-factor model

Before confirmatory factor analysis, we evaluated the three factors qualitatively. These factors are called personal application and relation to real world (PARRW), effort and sense making (ESM), and problem solving and learning (PSL). The latter was renamed problem solving self efficacy (PSSE) as described in Sec. III A.

TABLE I. The CLASS statements for the three factors: personal application and relation to real world, problem solving self efficacy (formerly called problem solving and learning), and effort and sense making according to Ref. [11]. Statement 25, marked with an asterisk, also loads on PSSE.

Factor	Number	Statement
PARRW	3.	I think about the physics I experience in everyday life.
	14.	I study physics to learn knowledge that will be useful in my life outside of school.
	25.*	I enjoy solving physics problems.
	28.	Learning physics changes my ideas about how the world works.
	30.	Reasoning skills used to understand physics can be helpful to me in my everyday life.
	37.	To understand physics, I sometimes think about my personal experiences and relate them to the topic being analyzed.
ESM	23.	In doing a physics problem, if my calculation gives a result very different from what I'd expect, I'd trust the calculation rather than going back through the problem.
	24.	In physics, it is important for me to make sense out of formulas before I can use them correctly.
	29.	To learn physics, I only need to memorize solutions to sample problems.
	32.	Spending a lot of time understanding where formulas come from is a waste of time.
PSSE (PSL)	5.	After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.
	21.	If I don't remember a particular equation needed to solve a problem on an exam, there's nothing much I can do (legally!) to come up with it.
	22.	If I want to apply a method used for solving one physics problem to another problem, the problems must involve very similar situations.
	34.	I can usually figure out a way to solve physics problems.
	40.	If I get stuck on a physics problem, there is no chance I'll figure it out on my own.

In order to investigate the relevance of the items with regard to the content aspect of the construct, we asked three additional independent expert raters (physics education researchers) to reassign the items to the three categories (PARRW, ESM, PSSE). The CLASS statements belonging to these factors are listed in Table I.

C. Statistical analysis of 3-factor model

The final model of Douglas *et al.* consists of 15 items [11]. As the simple 3-factor model did not meet the criteria for a sufficient model fit in CFA, the final model includes four adjustments to the model (1 crossloading, with item 25 loading on two factors, and 3 correlated errors between items 3 and 27, 25 and 37, and 21 and 34).

In order to obtain comparable results we replicated the methods reported by Douglas *et al.* We also used the software IBM AMOS with the same estimation method (asymptotic distribution free) and replaced missing values by the mean item value [11]. Overall only 0.38% of the answers on the 15 items were missing and the item with the maximum amount of missing answers was item 30 with 1.40% of missing values. As acceptance criteria for the model fit, we used meeting standard model fit criteria thresholds, as reported by Ref. [11]. Where this was not achieved, we settled for meeting the fit criteria for the final model in Ref. [11].

D. Think-aloud study

Ten undergraduate students were interviewed in a think-aloud study. Six of the students specialized in theoretical physics, two in physics, and two in meteorology or atmospheric sciences. Five students were male and five female. The most common specialization areas of the bachelor program in physical sciences at the University of Helsinki are physics, meteorology, and theoretical physics.

The students were asked to answer the 15 CLASS statements of the model of Ref. [11] on a 5-point Likert scale and justify their answers. The interviews were conducted in Finnish, but the statements were in English, as in the quantitative study. The interviews were recorded and the transcript coded for the factors per the definitions presented in Sec. III A by two persons. For answers which displayed reasoning consistent with more than one category, the primary one was determined.

Differences were reconciled with discussion to obtain a primary category.

III. RESULTS

A. Content aspect of validity

The primary assignment of the statements to the categories emerged from the discussion of the authors.

The distinction between the aspects of effort and sense making and problem solving and learning from these names was not immediately clear. We gave each factor more verbose definitions. The factor PARRW contained mainly statements about "relating physics to the world outside of formal physics education," of which "personal application and relation to real world" is a good summary. However, the presence of statement 25, "I enjoy solving physics questions," is difficult to reconcile with the other statements. It clearly measures enjoyment of problems, but does not necessarily relate to the real world.

The factor ESM contains questions that probe "using understanding related strategies in physics to learn or solve problems," for which we found "effort and sense making" an accurate summary. However, the questions encompassed by PSL were found to have qualities relating to "self-efficacy regarding problem solving situations in physics" rather than general problem solving. From here on, we will use the abbreviation PSSE (problem solving self efficacy) for this factor. Curiously, statement 25, which we found problematic for the definition of PARRW, also loads on this factor. This double loading has been problematic for Douglas *et al.* [11]. It seems plausible that enjoyment would be related to self-efficacy, but for the content aspect, it should not be part of the dimension self efficacy.

To conduct the secondary assignment, three expert raters were handed the 15 statements of the final model of Douglas *et al.* [11] together with the definitions of the three scales. They were asked to assign the statements to the three scales. They were also allowed to state that none of the statements was part of the scales. A summary of the expert rating is shown in Fig. 1. For these 15 items we calculated the interrater agreement coefficient Cohen's $\kappa_1 = 0.80$, $\kappa_2 = 0.81$, $\kappa_3 = 0.81$. Fleiss' kappa for all four raters was 0.745. According to Altmann [25] this can be interpreted overall as a good agreement.

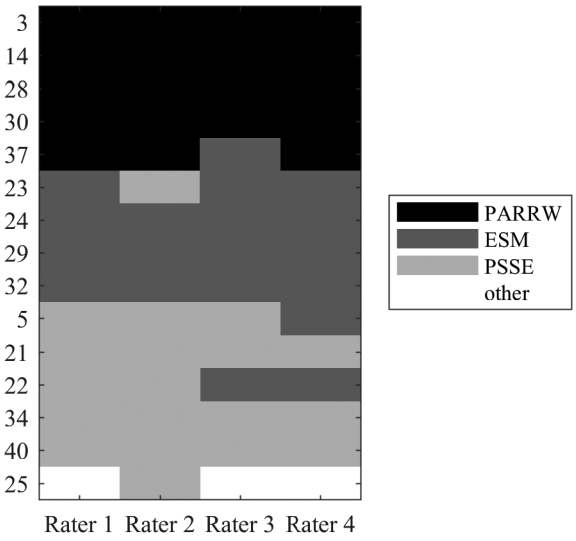

FIG. 1. Coding of the 15 statements included in the analysis. *X* axis: rater, *y* axis: question number.

TABLE II. Model fit indices for the CFA models by Douglas *et al.* [11] (model A–B) and the models based on the Finnish student sample (model 1–4)

Model	Analysis (Data)	Items (modifications)	χ^2 (df)	RMR	GFI	CFI	RMSEA	BIC
A	Douglas *et al.* (original)	15(0)	516.70(87)	0.071	0.928	0.675	0.051	766.3
B	Douglas *et al.* adj. (original)	15(4)	323.82(83)	0.058	0.955	0.818	0.039	603.7
1	Douglas *et al.* (Finnish)	15(0)	247.105(87)	0.076	0.921	0.686	0.054*	460.4
2	Douglas *et al.* adj. (Finnish)	15(4)	178.42(83)	0.065	0.943	0.813	0.042*	417.6
3	New model (Finnish)	14(0)	162.16(74)	0.069	0.945	0.806	0.043*	362.6
3a	New model crossload. (Finnish)	14(0)	162.08(73)	0.069	0.945	0.804	0.044*	368.47
4	New model adj. (Finnish)	14(4)	111.82(70)	0.057	0.962*	0.908*	0.031*	338.1
	Literature thresholds			relative	>0.95	>0.95	<0.05–0.08	Relative

*Fit criterion meets the standard of the literature (for CFI and RMSEA see Ref. [26] for GFI see Ref. [27]) or the values in Ref. [11]).

In one case the disagreement seemed substantial: It showed that statement 22 ("If I want to apply a method used for solving one physics problem to another problem, the problems must involve very similar situations.") might not be assigned only to PSSE but also to ESM. Two raters matched it to ESM. We did not deem the other disagreements substantial. E.g., there was disagreement with rater 2 regarding two statements: the authors doubted that "I enjoy solving physics problems" reflected any of the scales; hence we assigned it to the category other. So did raters 3 and 4. The second rater assigned this statement to PSSE, arguing that a student who had a high self efficacy would also enjoy solving physics problems. Also statements 5, 23, and 37 received one rating each that differed from the authors'.

B. Structural aspect of validity

In order to investigate the structural aspect of the construct we estimated four CFA models (1–4), which we compared to those of Douglas *et al.* (model A without and model B with *post hoc* adjustments; model B is the final model of Ref. [11]).

- Model 1: The simple 3-factor model structure of Douglas *et al.* [11] (model A), without *post hoc* adjustments. This model is included because it reflects the original hypothetical factor structure before adding the *post hoc* changes).
- Model 2: The previous model amended with the four adjustments reported by Douglas *et al.* (model B). This is model the final model structure of Douglas *et al.* [11].
- Model 3: The 3-factor model structure of Douglas *et al.* [11] (model A) with the exclusion of item 25 due to its lack of content validity regarding any of the constructs. We hypothesized that the lack of content validity also has impact on the empirical model fit.
- Model 3a: This model corresponds to the hypothesis, that item 22 might also be related to ESM and not just to PSL/PSSE. Consequently, the model is the same as model 3 but it includes a cross loading from item 22 on ESM.
- Model 4: Model 3 with four new adjustments (4 correlated errors). This model contains the same number of *post hoc* adjustments as the previous factor analysis. Modification indices identify the possible residual correlations. No cross loading was allowed to avoid within-item dimensionality concerning the three factors.

Table II shows the fit indices for the models estimated by Douglas *et al.* [11] (model A and B) as well as our estimations based on the Finnish student sample (model 1 to 4).

In order to accept or reject a model, we considered the fit criteria GFI, CFI, and RMSEA. For completeness, we report also RMR, χ^2 and BIC, but they will not be used here in order to reject the model. χ^2 is oversensitive for data sets of this size [26] and no absolute fit criteria exist for BIC and RMR [26]). In order to accept the model, every index needed to be in an acceptable range. We deemed the index to be in an acceptable range if it meets the literature thresholds reported in Table II. Indices not meeting this criterion were accepted, provided they were equal or better than those of the final model of Douglas *et al.* [11] (model B, Table II).

Based on these criteria, we evaluate the acceptability of our four models:

- Model 1: Using the same model specification as Douglas *et al.* [11] without the residual correlations and the crossloading, the model fit parameters do not fit those of model B, nor does model 1 meet the normative absolute threshold criteria. Only the RMSEA lies in the acceptable range.
- Model 2: The modifications used by Douglas *et al.* [11] (model B and model 2) do not lead to a substantial improvement. The RMSEA is still the only acceptable value.
- Model 3: In this model we excluded item 25 due to its lack of content validity regarding any of the constructs. While the fit parameters of the model are not

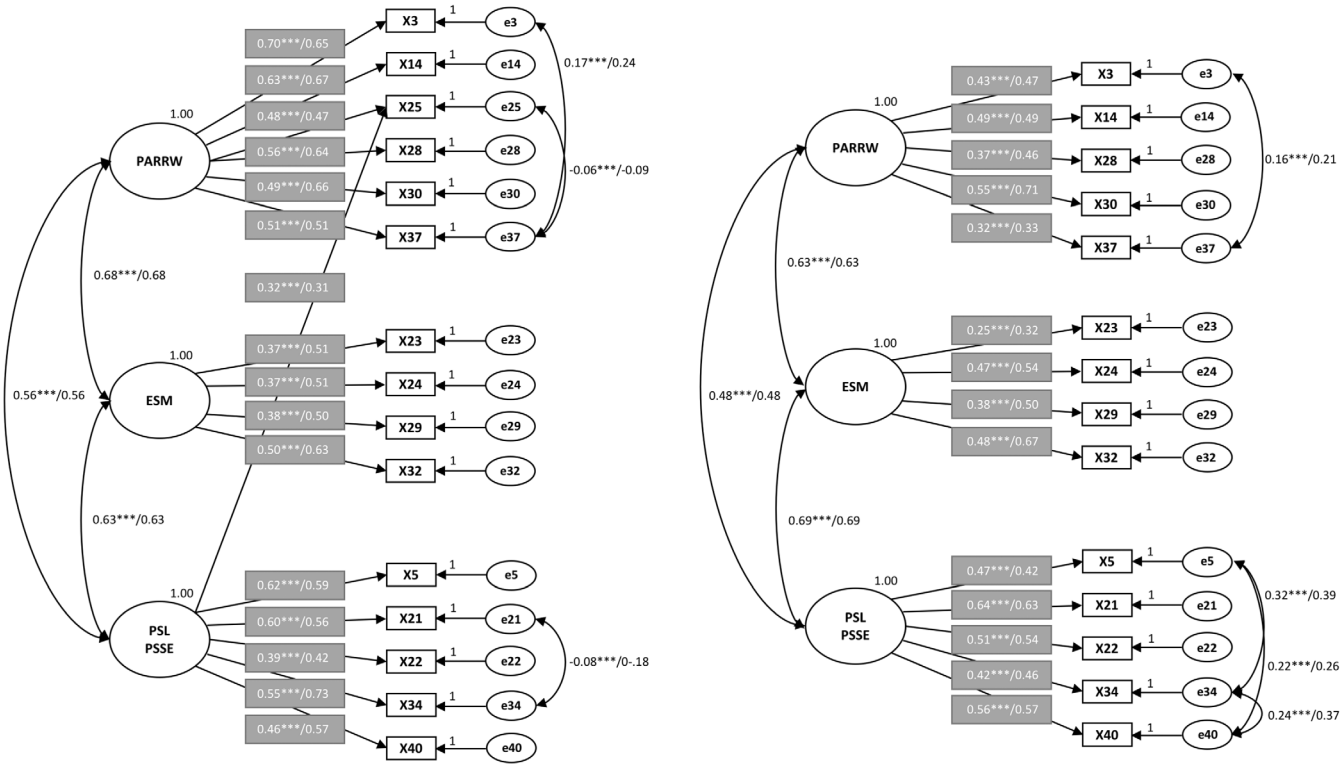

FIG. 2. Path diagrams for the CFA of Douglas *et al.* [11] (left) and the new model (right). The gray boxes contain the loadings also represented by the arrows. Two headed arrows represent the correlations of the constructs and residual correlations of the items. The variances of PARRW, ESM, and PSL and of the errors are set to one. The values behind the slashes are standardized.

acceptable, they are an improvement to those of the initial model 1.

- Model 3a: In this model we tested the hypothesis that emerged from the think-aloud study and the expert rating. It shows that there is no substantial loading from item 22 on the factor ESM (standardized and nonstandardized coefficient: 0.02) and no substantial improvement of any of the fit parameters compared to model 3.
- Model 4: Allowing the same number of *post hoc* adjustments as model B, we achieve fit parameters meeting literature standards (GFI and RMSEA) or those of the model B (CFI). The final model is similar to the model of Douglas *et al.* [11], but excludes item 25. The four residual correlations concern different items.

Consequently, model 4 provides the best description of our data. This leaves us with a model which is similar to the previously published model, but differs in both the number of items and the placement of the residual correlations (Fig. 2). The final model 4 does not have cross loadings.

In order to conduct measurement, the internal consistency needs to be in an acceptable range. As in Douglas *et al.* [11], we also used Cronbach's α as

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum(\widehat{\alpha_i}^2)}{\widehat{\alpha_X}^2}\right). \quad (1)$$

The number of items is denoted as k, $\widehat{\alpha_i}^2$ is the variance of item i, and $\widehat{\alpha_X}^2$ is the total variance of the scale. When comparing these values to those of Douglas *et al.* [11], we need to account for the exclusion of item 25, which leads to a lower Cronbach's α. The values for model 4 are $\alpha_{\mathrm{PARRW}} = 0.65, \alpha_{\mathrm{ESM}} = 0.48, \alpha_{\mathrm{PSSE}} = 0.71$, whereas those for the previously published model are $\alpha_{\mathrm{PARRW}} = 0.82$, $\alpha_{\mathrm{ESM}} = 0.61, \alpha_{\mathrm{PSSE}} = 0.73$.)

C. Think-aloud study

The interrater agreement for the ten student interviews was Cohen's $\kappa = 0.81$, which indicates high agreement [25]. The primary disagreements concerned items 5, 21, and 22, where both reviewers agreed some student answers contained elements of both ESM and PSSE, but differed on which was the primary. The differences were reconciled with discussion to obtain final categories for all answers.

The summary of the students' coded answers is presented in Fig. 3, which shows the fraction of student answers that conformed to the definitions developed earlier. For most items, the agreement is very good, but two items stand out.

To obtain an overview of how well the perspectives in student answers correspond to the ones assigned by the authors, we mapped the final coding of the student answers to the assigned categories. We calculated the agreement of student answers to the factors of model 4 using Cohen's κ,

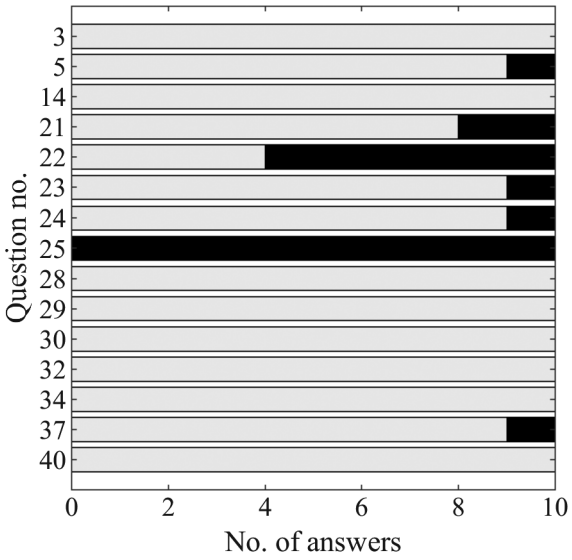

FIG. 3. The final coding of the 15 statements included in the analysis by the two raters. X axis: number of student answers, y axis: question number. Gray indicates answers coded to factors assigned by Douglas *et al.* [11], black indicates answers coded to other factors.

finding the average coefficient to be $\kappa = 0.87 \pm 0.10$. This again indicates high agreement. Cohen's κ was equal or above 0.8 for all students except one (for whom $\kappa = 0.62$), who had a higher number of answers coded as ESM.

While most student reasoning was as expected, the reasoning for some statements differed from the expected in ways we describe below.

1. Personal application and relation to real world

Statements 3, 14, 25, 28, 30, and 37 form the personal application and relation to real world factor in the analysis by Douglas *et al.* [11]. With the exception of statement 25, these form the same factor in our analysis.

The answers to all of these statements except statement 25 contain mostly PARRW justifications. The exception was one student's comments to statement 37 (*To understand physics, I sometimes think about my personal experiences and relate them to the topic being analyzed*), which was coded ESM:

"Personal experiences, does that mean your understanding of things or a specific experience? Well, I think understanding counts here. It is helpful for understanding to think how something actually works. Especially like mechanics problems, if you think it through, like which forces affect it and how it can go in real life. Not just writing equations and wondering whether it was correct. So I completely agree."

a. Statement 14 The coding of statement 14 (*I study physics to learn knowledge that will be useful in my life outside of school*) was straightforward and both raters agreed on all counts, but some interpretations of the students were surprising. One student of experimental physics was very clear on professional life not being involved in the word "school":

"Yes, well, I do not of course study for school or study physics for the grades, but to get a job and a profession and to be able to do it in my life after the studies."

However, the other students were less clear on whether school excluded academic research. A theoretical physics student disagreed with the statement:

"I would say partially disagree. Of course I study physics to get a profession out of it and it would kind of mean what they say here, life outside of school, but I do aim to get an academic career and the reason why I study physics is that I'm interested in it and the applications it can be used for, new inventions and stuff, but I don't really care about using it, like I'm not interested in what I think they're going for, like an everyday benefit from it."

The expertlike answer would be to agree with this statement. While some students agreed, several students disagreed, even though they reasoned (similarly to the agreeing students) that they aim for a career in physics. Also, a meteorology student partially agreed, but her reasoning was that the computer skills she has obtained might be useful to her in her spare time.

The confusion was not about whether the students understood the colloquial meaning of "outside of school," but how they related their education to their future career. The students viewed to a varying level their desired career as an extension or a direct consequence of their studies, which led to differences in their answers. To a direct question about whether working at the university is outside of school, one student said,

"Well when you put it like that, maybe if you're just an employee it's not school anymore, but now that I am a student here, I think [working at the university] is still like school, even when I work here during the summer."

This statement concisely summarizes the problem. When prompted, the student realizes that a practicing physicist in academia probably views their schooling and career as separate, but to him, having not reached this stage, they form a continuum. This interpretation is unlikely to arise when working with students who have not declared a major yet, or who have physics as a minor subject. Hence, items designed for a wide use may perform unexpectedly in populations consisting of physics majors, most of whom aim for careers in physics.

b. Statement 25 The open answers to statement 25 (*I enjoy solving physics problems*) bore little resemblance with the answers to other PARRW statements. The world outside formal physics education did not factor in. For example, one theoretical physics student answered,

"Well… Partly agree. I would not study physics if I did not enjoy it and if I did not find it interesting to solve physics problems, but if you think about like some calculation exercises then it depends on the exercise.

Some are really unpleasant and some are really fun. Usually it depends on like in the end how unpleasant and complicated mathematics you have to use in the solution. And on the other hand, sometimes problems are interesting in the way that solving them kind of helps you understand something big instead of just calculation methods."

All students referenced calculation exercises in their answers. Several students stated variations of "if I did not enjoy solving problems, I would not be studying physics" but they generally displayed the same ambivalence as the student quoted above: they found little enjoyment in long, tedious or difficult problems, but did enjoy problems that taught them something new, or problems that gave them a sense of accomplishment.

2. Effort and sense making

The statements numbered 23, 24, 29, and 32 make up the factor ESM. In the think-aloud study, statements 24, 29, and 32 contained wholly or mostly explanations that related to understanding. The students did sometimes make a distinction between the behavior they aspired to and the reality. For example, a student of theoretical physics commented on the statement 24 (*In physics, it is important for me to make sense out of formulas before I can use them correctly*):

"Yes, I would say that I at least mostly agree. You can often use formulas without completely understanding, like for example in quantum mechanics it is often easier to just calculate than to try to completely understand absolutely everything that the formulas mean but generally, like in classical mechanics where it's easier to see intuitively, being able to use formulas well and efficiently means that you can in practice understand the physics behind it and what it means in practice."

The statement 23 (*In doing a physics problem, if my calculation gives a result very different from what I'd expect, I'd trust the calculation rather than going back through the problem*) was more problematic. The explanations circled around there being a problem and needing to fix it, without a direct link to understanding. The same theoretical student as above said,

"This question sounds like… [you should not do] like this in physics. It is a rule of thumb that if your solution gives a numerical value which is significantly different to what you'd expect it usually means the answer is wrong."

There is an implicit connection to understanding the problem deeply and evaluating the answer based on other knowledge, and this answer was coded ESM, but the students' answers are operating on a higher level. They are making generalized statements about physics rather than their personal reasoning process.

3. Problem solving self-efficacy

Statements 5, 21, 22, 34, and 40 make up the PSSE factor. Of these, students used clear self-efficacy justification for statements 5, 21, 34, and 40. For example, one student said the following about statement 34 (*I can usually figure out a way to solve physics problems*):

"Yes, there have been only a few problems where I haven't come up with a solution. There have been some at times, especially at university level."

Many students also interpreted statement 21 [*If I don't remember a particular equation needed to solve a problem on an exam, there's nothing much I can do (legally!) to come up with it*] to be about self-efficacy. However, in the validation of CLASS items for predominantly Hispanic students, Sawtelle *et al.* regarded self-efficacy interpretations of statement 21 as misunderstandings [17]. Their correct interpretation was to state that equations can be derived, and answers referring to other ways of finding a solution were deemed to not align with experts. This question is regarded as being part of factors exploring conceptual understanding, not self-efficacy. We see this confusion in student answers: one meteorology student explicitly stated that she was unsure of the correct interpretation:

"I think it's a little unclear so I wouldn't know whether this 'come up with' means the equation or an answer to the whole problem." Her answer contained a reference to deriving the equation, but she emphasized other ways of answering questions. So did a student of theoretical physics: "Partially disagree. Of course you can derive things. But if then you can't figure it out you can try to continue with the problem or if you in principle know how to solve it you can just give a verbal explanation."

The answers to this question are likely depending on the type of exams students encounter. Two students explicitly said that they have little experience with exams where they are asked to remember a particular formula, as they are used to exams where crib sheets are allowed.

Statement 22 (*If I want to apply a method used for solving one physics problem to another problem, the problems must involve very similar situations*) was interpreted in different ways, either as a self-efficacy statement or about understanding related strategies. One physics student said,

"Well, here again the way I've felt is that I have had problems using some methods I've learned and applying them into some other topic."

On the other hand, a meteorology student thought about the methods as such:

"I'm not completely sure, I think it depends on what kind of method I use, like you can use some mathematical methods on many kinds of problems. […] Like differential equations have been used on many courses and in different situations and more generally if you think about statistical methods, they're applied in different fields, so I think they're different situations. So… but I think it maybe depends on the method, too. Well, it says "must involve" so I would say that I disagree. Because they don't have to be."

This answer was coded as ESM. In both cases, the students' answer aligned with the expertlike answer.

An expert ought to have high self-efficacy in applying a familiar method to a new problem, and an expert certainly should agree that in general in physics the same method can be applied in different situations. However, these justifications differ significantly. One student declined to answer, because he did not know what was meant by "method." When prompted about what he would answer in a written survey, he answered "neither agree nor disagree." He pondered both self-efficacy and ESM justifications, but as he did not use either to answer, these were not counted for either factor.

To summarize, the students mostly used explanations relating to self-efficacy for questions coded PSSE, but aspects of ESM appeared particularly for question 22, where they actually formed the majority. This result is similar to that of the expert rating.

IV. DISCUSSION

When studying the Finnish students' answers using the approach of Douglas *et al.* [11] and extending the analysis from a confirmatory factor analysis to an expert rating and student interviews, we found that the results obtained largely support each other. Factors similar to that presented in Ref. [11] (here called model B) are also present in the UH data, but certain differences run through the set of analyses conducted, leading us to modify the model. Many of these differences span both the content, substantive, and structural aspects of validity.

An important difference that emerged in all analyses was that the enjoyment of physics problem solving did not appear a part of any factor. While one expert rater assigned it to PSSE, his reasoning was not that enjoyment is a part of self-efficacy, but that enjoyment goes hand in hand with self-efficacy beliefs. Indeed, in model B, this item loads on two factors (PARRW and PSSE). However, this does not support the inclusion of item 25 in the factor PSSE when considering substantive validity. Neither does the item have much in common with the other items in PARRW. We therefore removed this item. This decision improved the results of the fit.

The broad phrasing of the CLASS statements that makes the survey usable in different contexts also makes the answers vulnerable to different interpretations. An example of this was seen with the interpretation of what "knowledge useful in my life outside of school" means. Surprisingly, students who have decided to pursue a degree in physics to get a career in physics may interpret this statement to mean knowledge which is not related to their future career, which compromises substantive validity. However, despite the interviews showing different interpretations of this item, in the quantitative analysis, this statement still mapped on the PARRW factor. Disagreeing with this expertlike answer while showing expertlike responses in other questions is hence unlikely to be very common in the whole sample.

More problematic were the interpretations of statement 22. The expert opinions were split evenly between assigning this statement to PSSE and ESM. Three of the students interpreted this statement to be clearly about self-efficacy, while six others used justifications that were better mapped to sense making. In addition, one student declined to answer, as he thought the question was too broad. Still, in the quantitative analysis this question maps to PSSE, and a cross loading on ESM did not improve the fit and was close to zero. Hence, we see a problem with the content validity, but this could not be confirmed in the factor structure. The problem does not seem to be generalizable from a statistical point of view.

As validity is a property of the test score interpretation rather than a property of the test itself [20], it seems useful to ask which kind of interpretations we consider valid given in the context of testing Finnish physics students. Given some problematic individual interpretations of the statements, individual diagnostic assessment on the three factors should not be implemented. However, as the dimensionality seems transferable in a statistical sense and as the alternative interpretations are limited in amount, the scores for PRRW, ESM, and PSL might be used in order to obtain population values for a larger group of physics students. Given a low mean value (e.g., in PARRW) the lecturers or teachers could then decide whether they intervene (in this example, by including physics problems which link physics to the real world).

Our results are obtained using a subset of the CLASS items, leading to an abbreviated CLASS. However, the original CLASS contains many statements which are similar to each other. For example, the ESM statement "In physics, it is important for me to make sense out of formulas before I can use them correctly" is related to the omitted statement "When I solve a physics problem, I locate an equation that uses the variables given in the problem and plug in the values." These statements are not the same, but the first statement probes the same idea as the second: how the student understands the role of formulas in physics. Excluding one therefore does not remove the idea of formula use from the test, even if the phrase may not be identical.

However, there is one important aspect which is not covered at all: questions that relate to learning preferences. For example, the statements 12 and 19, "I cannot learn physics if the teacher does not explain things well in class" and "To understand physics I discuss it with friends and other students" do not have counterparts in the statements included in our CFA. These statements are also not included in the original 8-factor structure, but are scored in the overall CLASS score [8]. Also statement 25, which we removed in our analysis, could be seen to fall into this category of personal preference, but previous factor analyses use it [8,11]. Using the abbreviated CLASS leads to a loss of these aspects.

While the external aspect of validity of CLASS has been studied previously by, e.g., correlating CLASS results to conceptual learning measured by the Force Concept Inventory [4,5], Brief Electric and Magnetic Assessment [5], and the Force and Motion Conceptual Evaluation [2], this is beyond the scope of our study. Previously, Hendolin has correlated the CLASS factors from model B to factors in optimal learning moments (OLM), finding evidence for a statistically significant correlation between the OLM Skill and OLM Interest factors and those of model B [13], but as the CLASS factors used in that study are slightly different than those obtained by us, there is a need for further validation.

To our knowledge, CLASS has not been used to study systematically the development of expertlike attitudes in Finnish physics students. As the development of expertlike attitudes is an important goal in educating future physicists, the validation of CLASS for Finnish students enables further studies where the development of expertlike attitudes of Finnish university students can be reliably compared. For example, a recent paper shows that Finnish female physics students have lower self-efficacy than their male counterparts [28], which might also show in the CLASS results.

V. CONCLUSIONS

We have studied the validity of the 3-factor structure of the CLASS instrument by confirmatory factor analysis, expert rating, and student interviews. Confirmatory factor analysis is a standard tool in assessing test validity, and our results show that the factor structure obtained from student answers in a U.S. institution [11] is similar to that obtained at the University of Helsinki. For the factor structure, we found a better fit by dropping one item (reducing the number of items from 15 to 14) and assigning different correlated errors. This also leads to an improvement in structural validity of the factor model, as the dropped item had a problematic double loading in the previous model.

The factor analysis results support the use of CLASS in varied student populations. Also, student interviews and expert ratings mostly agreed with the factor analysis. However, we did uncover some important discrepancies.

CLASS statements are intentionally phrased broadly for the instrument to be useful in many different contexts [8]. When extending the use of instruments to new (non-English speaking) populations, a lot of care is usually taken to ensure that the students understand the questions. We found that while our students had no problem parsing the CLASS statements, the justifications they used in their answers were sometimes difficult to assign. While the overall agreement was high, reasoning was not consistent for some items. In particular, assignments of item 22 split both the experts and the students. We also observed self-efficacy interpretations of item 21, which have previously been considered misconceptions [17]. In many cases, the students' answer had elements of several factors.

A new misconception that we observed was that some high-achieving students not linking their everyday experience to their studies, as their physics studies were motivated by an interest in theoretical physics or abstract topics. These kinds of inconsistencies reduce the content validity of the instrument, but are not visible in a confirmatory factor analysis. This also highlights the importance of validating items not only by experts, but by students who represent a similar demographic as those tested. However, these misinterpretations concern only a few of the studied statements, and some misinterpretations are similar to those observed previously. Hence, it seems likely that the factor model for the Finnish students is equally reliable as the previous one for its respective population.

Validation should not be regarded as a process that will be ever completely finished [20]. When the use of instruments is extended to a new population, it should be rigorously studied. We found that the previous CLASS factor analysis was mostly generalizable to a new population. However, students interpret the broad statements based on their own experiences and expectations of physics studies and prospective careers, and surprising interpretations can emerge. Because of its wide use, CLASS is an important instrument for surveying students' attitudes. Another important feature is the authenticity and relevance of the CLASS statements. It is important to note that the CLASS statements were assigned to the scales after their construction. This makes it more difficult to construct scales containing distinct sets of items. However, these statements are considered highly relevant by experts, which increases the authenticity and relevance concerning successful learning of physics. In our opinion, this is an important reason to continue using this instrument. We demonstrate the need to study answering patterns and student interpretations in different student populations, and to explore the underlying constructs further.

ACKNOWLEDGMENTS

The authors want to thank Elina Palmgren for coding the student interviews and for insightful discussions, Kimmo Kulmala for help with the student interviews, Joost Massolt, Tanja Mutschler, and Elina Palmgren for assigning factor definitions, Professor Ian Bearden for discussions and helpful comments and the three anonymous reviewers for their constructive and thoughtful comments that improved this paper.

[1] K. E. Gray, W. K. Adams, C. E. Wieman, and K. K. Perkins, Students know what physicists believe, but they don't agree: A study using the class survey, Phys. Rev. ST Phys. Educ. Res. **4**, 020106 (2008).

[2] K. Perkins, W. Adams, S. Pollock, N. Finkelstein, and C. Wieman, Correlating student beliefs with student learning using the Colorado Learning Attitudes about, Science Survey, AIP Conf. Proc. **790**, 61 (2005).

[3] G. Kortemeyer, Correlations between student discussion behavior, attitudes, and learning, Phys. Rev. ST Phys. Educ. Res. **3**, 010101 (2007).

[4] L. Ding, Verification of causal influences of reasoning skills and epistemology on physics conceptual learning, Phys. Rev. ST Phys. Educ. Res. **10**, 023101 (2014).

[5] M. J. Cahill, M. A. McDaniel, R. F. Frey, K. M. Hynes, M. Repice, J. Zhao, and R. Trousil, Understanding the relationship between student attitudes and student learning, Phys. Rev. Phys. Educ. Res. **14**, 010107 (2018).

[6] I. Halloun, Views about science and physics achievement: The VASS story, AIP Conf. Proc. **399**, 605 (1997).

[7] E. F. Redish, J. M. Saul, and R. N. Steinberg, Student expectations in introductory physics, Am. J. Phys. **66**, 212 (1998).

[8] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).

[9] W. K. Adams, C. E. Wieman, K. K. Perkins, and J. Barbera, Modifying and validating the Colorado Learning Attitudes about Science Survey for use in chemistry, J. Chem. Educ. **85**, 1435 (2008).

[10] K. Semsar, J. K. Knight, G. Birol, and M. K. Smith, The Colorado Learning Attitudes about Science Survey (class) for use in biology, CBE Life Sci. Educ. **10**, 268 (2011).

[11] K. A. Douglas, M. S. Yale, D. E. Bennett, M. P. Haugan, and L. A. Bryan, Evaluation of Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **10**, 020128 (2014).

[12] M. J. Cahill, K. M. Hynes, R. Trousil, L. A. Brooks, M. A. McDaniel, M. Repice, J. Zhao, and R. F. Frey, Multiyear, multi-instructor evaluation of a large-class interactive-engagement curriculum, Phys. Rev. ST Phys. Educ. Res. **10**, 020101 (2014).

[13] I. Hendolin, Exploring optimal learning moments at tutorial sessions, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA*, edited by D. L. Jones, L. Ding, and A. Traxler (AIP, New York, 2016), pp. 144–147, https://doi.org/10.1119/perc.2016.pr.031.

[14] P. Zhang and L. Ding, Large-scale survey of chinese precollege students' epistemological beliefs about physics: A progression or a regression?, Phys. Rev. ST Phys. Educ. Res. **9**, 010110 (2013).

[15] H. Alhadlaq, F. Alshaya, S. Alabdulkareem, K. K. Perkins, W. K. Adams, and C. E. Wieman, Measuring students beliefs about physics in Saudi Arabia, AIP Conf. Proc. **1179**, 69 (2009).

[16] J. de la Garza and H. Alarcon, Assessing students' attitudes in a college physics course in Mexico, AIP Conf. Proc. **1289**, 129 (2010).

[17] V. Sawtelle, E. Brewe, and L. Kramer, Validation study of the Colorado Learning Attitudes about Science Survey at a hispanic-serving institution, Phys. Rev. ST Phys. Educ. Res. **5**, 023101 (2009).

[18] J. Vettenranta, J. Välijärvi, A. Ahonen, J. Hautamäki, J. Hiltunen, K. Leino, S. Lähteinen, K. Nissinen, V. Nissinen, E. Puhakka, J. Rautopuro, and M.-P. Vainikainen, PISA 15 Ensituloksia. Huipulla pudotuksesta huolimatta, Tech. Rep. 41, Opetus ja kulttuuriministeriö, 2016.

[19] J. Lavonen and S. Laaksonen, Context of teaching and learning school science in Finland: Reflections on pisa 2006 results, J. Res. Sci. Teach. **46**, 922 (2009).

[20] S. Messick, Validity of psychological assessment, Am. Psychol. **50**, 741 (1995).

[21] R. L. Linn, Validation of the uses and interpretations of results of state assessment and accountability systems, in *Large-Scale Assessment Programs for All Students: Validity, Technical Adequacy, and Implementation*, edited by T. M. Haladyna and G. Tindal (L. Erlbaum, Mahwah, NJ, 2002), p. 523.

[22] P. Baghaei and M. Tabatabaee Yazdi, The logic of latent variable analysis as validity evidence in psychological measurement, Open Psychol. J. **9**, 168 (2016).

[23] R. B. Kline, *Principles and Practice of Structural Equation Modeling*, 3rd ed. (Guilford Press, New York, 2011).

[24] D. L. Jackson, Structural Equation Modeling: A adding missing-data-relevant variables to FIML-based structural equation models, Struct. Eq. Modeling **10**, 128 (2003).

[25] D. Altmann, *Practical Statistics for Medical Research* (Chapman and Hall, London, 1991).

[26] T. A. Brown, *Confirmatory Factor Analysis for Applied Research* (Guilford Press, New York, London, 2006), p. 475.

[27] J. N. Miles and M. Shevlin, Effects of sample size, model specification and factor loadings on the GFI in confirmatory factor analysis, Personality Indiv. Diff. **25**, 85 (1998).

[28] R. S. Barthelemy and A. V. Knaub, Gendered motivations and aspirations of university physics students in Finland, Phys. Rev. Phys. Educ. Res. **16**, 010133 (2020).